f# for data analysis

Артём Акуляков

Я

- dotnet, python, js, go
- $f(\lambda)$, data analysis, ml
- fintech стартап, senior engineer



data analysis

data analysis

- math and computer science
- СОСТОИТ
 - очистка
 - трансформация
 - дополнение
 - фильтрация
 - моделирование
 - кластеризация
 - поиск корреляций
 - проверка гипотез
 - ...
- ... rocket sience

data analysis не всегда rocket sience

- проект на поддержку без описания структуры данных
- расследование по логам
- сложная аналитика
- ...

житейский data analysis

f#

f# разработчики самые счастливые

f#

f#

- functional-first programming language
- компилируемый & интерпретируемый
- dotnet
- linux, osx, win, +
- монады, матан и вся страшная жесть
- хорош для data analysis

data analysis?

data analysis

- 1. доступ к данным
- 2. визуализация
- 3. манипулирование данными
- 4. интеграция(?)

1 доступ к данным

1 доступ к данным

FSharp.Data & FSharp.Data.TypeProviders

- sql db
- web & files
 - json
 - xml
 - CSV
 - html
- world bank
- twitter
- ...

```
open FSharp.Data

type GitHub = JsonProvider<"https://api.github.com/repos/dotnet/coreclr/issues">

let topRecentlyUpdatedIssues =
   GitHub.GetSamples()
   |> Seq.sortBy (fun x -> x.UpdatedAt)
   |> Seq.truncate 5
```

```
open FSharp.Data
```

```
type GitHub = JsonProvider<"https://api.github.com/repos/dotnet/coreclr/issues">
let topRecentlyUpdatedIssues =
   GitHub.GetSamples()
   |> Seq.sortBy (fun x -> x.UpdatedAt)
   |> Seq.truncate 5
```

```
open FSharp.Data

type GitHub = JsonProvider<"https://api.github.com/repos/dotnet/coreclr/issues">

let topRecentlyUpdatedIssues =
   GitHub.GetSamples()

|> Seq.sortBy (fun x -> x.UpdatedAt)
|> Seq.truncate 5
```

```
type GitHub = JsonProvider<"https://api.github.com/repos/dotnet/coreclr/issues">
     seq<JsonProvider<...>.Root>
     let topRecentlyUpdatedIssues =
         GitHub.GetSamples()
            Seq.sortBy (fun x \rightarrow x.)
            Seq.truncate 5
                                      Assignee
                                                                                   property J
                                      ✗ Assignees
                                                                                   ee: Option
     for issue in topRecentlyUpdate > AuthorAssociation
23
         printfn "#%d %s" issue.Id ▶ Body

    ClosedAt

                                      Comments
                                      ✗ CommentsUrl
                                      ✗ CreatedAt

№ EventsUrl

       id
            event id date created
                     10/31/2014 10:5 ► HtmlUrl
   -> 13
                     10/31/2014 1:01 5 Id
   -> 14
                     10/31/2014 6:11 № JsonValue
   -> 16
    -> 18
                     10/31/2014 8:05:21 PM
                     10/21/2014 0:10:04 DM
```

2 визуализация

2 визуализация

- xplot
- FSharp.Charting

```
open FSharp.Charting
open System

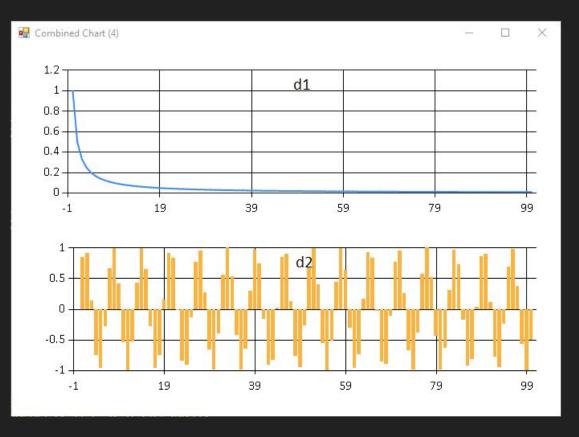
let d1 = [for x in 0 .. 100 -> (x, 1.0 / (float x + 1.) )]
let d2 = [for x in 0 .. 100 -> (x, Math.Sin(float(x)))]

Chart.Rows [
    Chart.Line(d1,Name="d1",Title="d1")
    Chart.Column(d2,Name="d2",Title="d2")
]
```

```
open FSharp.Charting
open System

let d1 = [for x in 0 .. 100 -> (x, 1.0 / (float x + 1.) )]
let d2 = [for x in 0 .. 100 -> (x, Math.Sin(float(x)))]

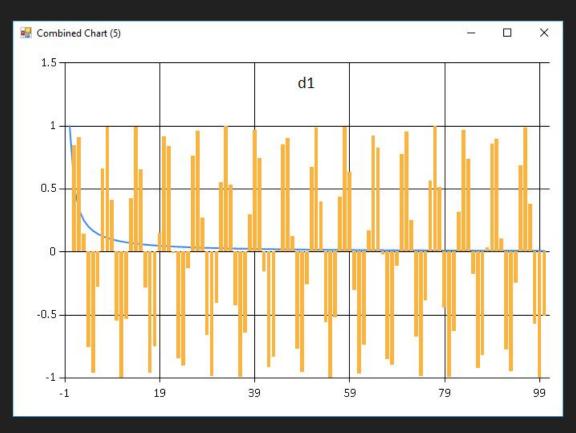
Chart.Rows [
    Chart.Line(d1,Name="d1",Title="d1")
    Chart.Column(d2,Name="d2",Title="d2")
]
```



```
open FSharp.Charting
open System

let d1 = [for x in 0 .. 100 -> (x, 1.0 / (float x + 1.) )]
let d2 = [for x in 0 .. 100 -> (x, Math.Sin(float(x)))]

Chart.Combine [
    Chart.Line(d1,Name="d1",Title="d1")
    Chart.Column(d2,Name="d2",Title="d2")
]
```



3 манипулирование данными

3 манипулирование данными

- repl & $f(\lambda)$

3 манипулирование данными

- repl & $f(\lambda)$
- Deedle

- аналог pandas из мира python
- недо-Excel на стероидах
- приправа из статистики

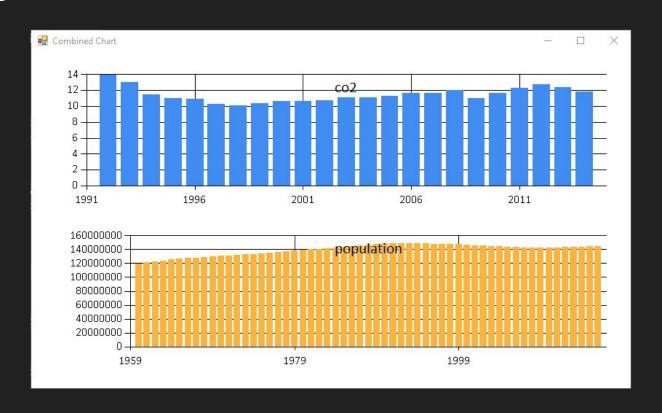
- Series
- Frames
- Stats

```
open FSharp.Data
let WorldBank = WorldBankData.GetDataContext()
let co2Indicator =
   WorldBank
       .Countries. ``Russian Federation``
       .Indicators. ``CO2 emissions (metric tons per capita) ``
let populationIndicator =
   WorldBank
       .Countries. ``Russian Federation`
       .Indicators. ``Population, total``
```

```
open Deedle
open FSharp.Data
let WorldBank = ...
let co2Series = co2Indicator |> Series.ofObservations
let populationSeries = populationIndicator |> Series.ofObservations
let frame = Frame(["co2"; "population"], [co2Series; populationSeries]) |>Frame.dropSparseRows
frame?totalCo2 <- frame?co2 * frame?population</pre>
let avarageCo2 = frame?totalCo2 |> Stats.median
```

```
open Deedle
open FSharp.Data
let WorldBank = ...
let co2Series = co2Indicator |> Series.ofObservations
let populationSeries = populationIndicator |> Series.ofObservations
let frame = Frame(["co2"; "population"], [co2Series; populationSeries]) |>Frame.dropSparseRows
frame?totalCo2 <- frame?co2 * frame?population</pre>
let avarageCo2 = frame?totalCo2 |> Stats.median
```

```
open Deedle
open FSharp.Data
let WorldBank = ...
let co2Series = co2Indicator |> Series.ofObservations
let populationSeries = populationIndicator |> Series.ofObservations
let frame = Frame(["co2"; "population"], [co2Series; populationSeries]) |>Frame.dropSparseRows
frame?totalCo2 <- frame?co2 * frame?population</pre>
let avarageCo2 = frame?totalCo2 |> Stats.median
```



Deedle

```
open Deedle
open FSharp.Data
let WorldBank = ...
let co2Series = co2Indicator |> Series.ofObservations
let populationSeries = populationIndicator |> Series.ofObservations
let frame = Frame(["co2"; "population"], [co2Series; populationSeries]) |>Frame.dropSparseRows
frame?totalCo2 <- frame?co2 * frame?population</pre>
let avarageCo2 = frame?totalCo2 |> Stats.median
```

Deedle

```
open Deedle
open FSharp.Data
let WorldBank = ...
let co2Series = co2Indicator |> Series.ofObservations
let populationSeries = populationIndicator |> Series.ofObservations
let frame = Frame(["co2"; "population"], [co2Series; populationSeries]) |>Frame.dropSparseRows
frame?totalCo2 <- frame?co2 * frame?population</pre>
let avarageCo2 = frame?totalCo2 |> Stats.median
```

Deedle

```
open Deedle
open FSharp.Data
let WorldBank = ...
let co2Series = co2Indicator |> Series.ofObservations
let populationSeries = populationIndicator |> Series.ofObservations
let frame = Frame(["co2"; "population"], [co2Series; populationSeries]) |>Frame.dropSparseRows
frame?totalCo2 <- frame?co2 * frame?population</pre>
let avarageCo2 = frame?totalCo2 |> Stats.median
```

3 манипулирование данными

- repl & $f(\lambda)$
- Deedle
- Math.Net

- матричные вычисления
- статистика
- решение систем уравнений
- регрессия
- ...

```
open MathNet.Numerics.Statistics
...
let co2Values = frame?co2 |> Series.values
let populationValues = frame?population |> Series.values
let coef = Correlation.Spearman(populationValues, co2Values)
```

```
open MathNet.Numerics.Statistics
...
let co2Values = frame?co2 |> Series.values
let populationValues = frame?population |> Series.values
let coef = Correlation.Spearman(populationValues, co2Values)
```

```
open MathNet.Numerics.Statistics

...
let co2Values = frame?co2 |> Series.values
let populationValues = frame?population |> Series.values

let coef = Correlation.Spearman (populationValues, co2Values)
```

```
open MathNet.Numerics.Statistics

...
let co2Values = frame?co2 |> Series.values
let populationValues = frame?population |> Series.values

let coef = Correlation.Spearman(populationValues, co2Values)
?> -0.2756916996
```

3 манипулирование данными

- repl & $f(\lambda)$
- Deedle
- Math.Net
- R Provider

- удобно получать и обрабатывать данные с f#
- удобно делать расчеты и визуализацию с R

```
open RDotNet
open RProvider

let rng = Random()
let rand () = rng.NextDouble()

let X1s = [ for i in 0 .. 9 -> 10. * rand () ]
let X2s = [ for i in 0 .. 9 -> 5. * rand () ]
let Ys = [ for i in 0 .. 9 -> 5. + 3. * X1s.[i] - 2. * X2s.[i] + rand () ]

R.plot(Ys)
```

```
open RDotNet
open RProvider

let rng = Random()
let rand () = rng.NextDouble()

let X1s = [ for i in 0 .. 9 -> 10. * rand () |
let X2s = [ for i in 0 .. 9 -> 5. * rand () |
let Ys = [ for i in 0 .. 9 -> 5. + 3. * X1s.[i] - 2. * X2s.[i] + rand () ]

R.plot(Ys)
```

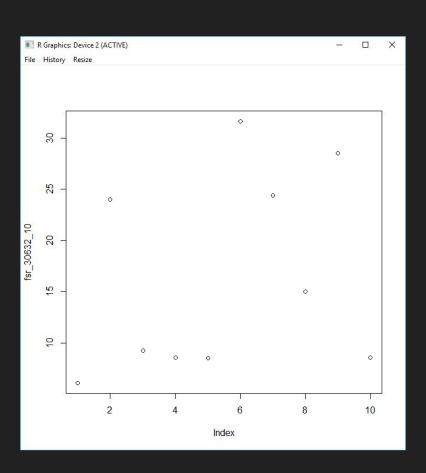
```
open RDotNet
open RProvider
let rng = Random()
let rand () = rng.NextDouble()
let X1s = [ for i in 0 .. 9 -> 10. * rand () ]
let X2s = [ for i in 0 .. 9 -> 5. * rand () ]
let Ys = [ for i in 0 .. 9 -> 5. + 3. * X1s.[i] - 2. * X2s.[i] + rand () ]
R.plot(Ys)
```

```
open RDotNet
open RProvider

let rng = Random()
let rand () = rng.NextDouble()

let X1s = [ for i in 0 .. 9 -> 10. * rand () ]
let X2s = [ for i in 0 .. 9 -> 5. * rand () ]
let Ys = [ for i in 0 .. 9 -> 5. + 3. * X1s.[i] - 2. * X2s.[i] + rand () ]

R.plot(Ys)
```



```
. . .
```

```
let ds = namedParams ["Y", box Ys;"X1", box X1s;"X2", box X2s;] |> R.data_frame
let result = R.lm(formula = "Y~X1+X2", data = ds)
```

. . .

```
let ds = namedParams ["Y", box Ys;"X1", box X1s;"X2", box X2s;] |> R.data_frame
let result = R.lm(formula = "Y~X1+X2", data = ds)
```

```
• • •
```

```
let ds = namedParams ["Y", box Ys;"X1", box X1s;"X2", box X2s;] |> R.data_frame
let result = R.lm(formula = "Y~X1+X2", data = ds)
```

3 манипулирование данными

- repl & $f(\lambda)$
- Deedle
- Math.Net
- R Provider
- {m}brace

- azure
- aws(?)

```
open MBrace.Core
let sourceValues = [|for x in 0 .. 16 \rightarrow [|for y in 0 .. 2000 \rightarrow rand()|]|]
let cluster = ThespianCluster.InitOnCurrentMachine(4, ...)
   [|for x in sourceValues -> cloud { return x |> Array.sum}|]
   > Cloud.Parallel
   > cluster.Run
   |> Array.sum
cluster.KillAllWorkers()
```

```
open MBrace.Core
let sourceValues = [|for x in 0 .. 16 \rightarrow [|for y in 0 .. 2000 \rightarrow rand()|]|]
let cluster = ThespianCluster.InitOnCurrentMachine(4, ...)
   [|for x in sourceValues -> cloud { return x |> Array.sum}|]
   > Cloud.Parallel
   > cluster.Run
   > Array.sum
cluster.KillAllWorkers()
```

```
open MBrace.Core
let sourceValues = [|for x in 0 .. 16 \rightarrow [|for y in 0 .. 2000 \rightarrow rand()|]|]
let cluster = ThespianCluster.InitOnCurrentMachine(4, ...)
   [|for x in sourceValues -> cloud { return x |> Array.sum}|]
   > Cloud.Parallel
   > cluster.Run
   |> Array.sum
cluster.KillAllWorkers()
```

```
open MBrace.Core
let sourceValues = [|for x in 0 .. 16 \rightarrow [|for y in 0 .. 2000 \rightarrow rand()|]|]
let cluster = ThespianCluster.InitOnCurrentMachine(4, ...)
   [|for x in sourceValues -> cloud { return x |> Array.sum}|]
   > Cloud.Parallel
   > cluster.Run
   > Array.sum
cluster.KillAllWorkers()
```

```
open MBrace.Core
let sourceValues = [|for x in 0 .. 16 \rightarrow [|for y in 0 .. 2000 \rightarrow rand()|]|]
let cluster = ThespianCluster.InitOnCurrentMachine(4, ...)
   [|for x in sourceValues -> cloud { return x |> Array.sum}|]
   > Cloud.Parallel
   > cluster.Run
   |> Array.sum
cluster.KillAllWorkers()
```

```
open MBrace.Core
let sourceValues = [|for x in 0 .. 16 \rightarrow [|for y in 0 .. 2000 \rightarrow rand()|]|]
let cluster = ThespianCluster.InitOnCurrentMachine(4, ...)
   [|for x in sourceValues -> cloud { return x |> Array.sum}|]
   > Cloud.Parallel
   > cluster.Run
   |> Array.sum
cluster.KillAllWorkers()
```

```
open MBrace.Core
let sourceValues = [|for x in 0 .. 16 \rightarrow [|for y in 0 .. 2000 \rightarrow rand()|]|]
let cluster = ThespianCluster.InitOnCurrentMachine(4, ...)
   [|for x in sourceValues -> cloud { return x |> Array.sum}|]
   > Cloud.Parallel
   |> cluster.Run
   |> Array.sum
cluster.KillAllWorkers()
```

```
mbrace.fsx • • deedle-mathnet-chartina.fsx
                                                                                         charting.fsx •
EXPLORER
 MBrace.Thespian Slave Node [pid:19900, port:59714]
                                                                                                                             ConsoleLogger(), logLevel = LogLevel.1
INFO: Initialized controller actor:
        MBrace.Thespian Slave Node [pid:18052, port:59715]
           MBrace.Thespian Slave Node [pid:17064, port:59713]
INFO :
          INFO: Thespian listening to WORK:59886 on AppDomain 'c5155d61-5817-48e3-be67-f29541c3a043'.
          INFO : Starting cloud process 'a8230c57-6ca5-4c35-b2fb-d9cdde86966d' of type 'float []'.
INFO :
INFO: TNF, INFO: Starting Cloud.Parallel<float> workflow 0e5108ef-99b1-4d0b-8332-298b753ae03b of 17 children.
INFO : INFINFO : Completed work item '6087c8a0-9f62-417a-a96b-351bc8a97f67' after 00:00:00.6530842
INFO : INF INFO : Concurrent work ite
                                     MBrace.Thespian Slave Node [pid:16084, port:59716]
                                                                                                                                                       INFO : INF INFO : Dequeued work item
INFO : TNE INFO : Concurrent work it INFO : Process Id: 16084
INFO : TNF INFO : Completed work itemINFO : Thespian initialized with listener WORK:59716
INFO : INFINFO : Concurrent work ite INFO : Initialized controller actor:
INFO : TNF INFO : Dequeued work item
                                            actor://controller.WorkerControllerMsg
      INFINFO : Concurrent work ite
                                            utcp://work:59716/controller
INFO : INF INFO : Completed work item
                                            mailbox://controller/.
n=null a3a INFO : Concurrent work ite INFO : Responding to spawning process.
INFO: TNF INFO: Dequeued work item INFO: Subscribing worker to runtime hosted at 'mbrace://WORK:59269'.
INFO : INFO : Concurrent work ite INFO : Initializing worker store logger.
INFO : n=n INFO : Completed work item INFO : Initializing AppDomain pool evaluator.
INFO : INFO : Concurrent work ite INFO : Creating worker agent.
INFO: INFIINFO: Dequeued work item INFO: Subscribed to cluster with heartbeat interval 00:00:01
INFO : TNF INFO : Concurrent work ite INFO : Starting Worker.
INFO : TNF,INFO : Completed work itemINFO : Dequeued work item '4a0fa06c-35cb-43a1-9b85-b5b948eddb19'.
INFO : INFO : Concurrent work iteINFO : Concurrent work item count increased to 1/20.
INFO : INFINFO : Dequeued work item INFO : Downloading 'FSharp.Compiler.Interactive.Settings, Version=4.4.1.0, Culture=neutral, PublicKeyToken=b03f5f7f11d50
INFO : INF INFO : Concurrent work iteasa'
INFO: TNE INFO: Completed work item_INFO: Downloading 'FSharp.Charting, Version=0.90.14.0, Culture=neutral, PublicKeyToken=null'
       TIME INFO: Concurrent work it INFO: Downloading 'FSI-ASSEMBLY 06633297-ed7e-4626-9461-32114d953896_2, Version=0.0.0.0, Culture=neutral, PublicKeyToke
 SysINFINFO : Dequeued work item n=null'
       INF|INFO : Concurrent work iteINFO : Dequeued work item '4b6568ba-7759-4<u>1c7-9d69-5083263f8c8c'.</u>
          INFO : Completed work itemINFO : Concurrent work item count increased to 2/20.
         INFO : Concurrent work ite<sub>INFO</sub> : Dequeued work item '89247bf6-71eb-49ed-b879-caa457ba6b75'.
 System
                                    INFO: Concurrent work item count increased to 3/20.
                                    INFO : Downloading data dependency 'Double[][] sourceValues@'.
 TaskParallelLibrary
                                    INFO : Downloading data dependency 'MBrace.Core.Cloud'1[System.Double[]] x@1-4'.
 ▶ Thespian
                                    INFO : Downloading data dependency 'MBrace.Core.Cloud'1[System.Double][] x@1-5'.
                                                                                                                                                                  .57 MiE
                                    INFO : Initializing Application Domain '94b2bf36-460c-4a15-a3b7-ebb7e85946ec'.
 ▶ Vagabond
                                                                                                                                                                 31bf385
                                    INFO : Thespian listening to WORK:59936 on AppDomain '94b2bf36-460c-4a15-a3b7-ebb7e85946ec'.

    Zlib.Portable

                                      INFO: Uploading 'FSharp.Compiler.Interactive.Settings, Version=4.4.1.0, Culture=neutral, PublicKeyToken=b03f5f7f1
paket-files
```

```
open MBrace.Core
let sourceValues = [|for x in 0 .. 16 \rightarrow [|for y in 0 .. 2000 \rightarrow rand()|]|]
let cluster = ThespianCluster.InitOnCurrentMachine(4, ...)
   [|for x in sourceValues -> cloud { return x |> Array.sum}|]
   > Cloud.Parallel
   > cluster.Run
   |> Array.sum
cluster.KillAllWorkers()
```

```
open MBrace.Core
let sourceValues = [|for x in 0 .. 16 \rightarrow [|for y in 0 .. 2000 \rightarrow rand()|]|]
let cluster = ThespianCluster.InitOnCurrentMachine(4, ...)
   [|for x in sourceValues -> cloud { return x |> Array.sum}|]
   > Cloud.Parallel
   > cluster.Run
   |> Array.sum
```

cluster.KillAllWorkers()

```
open MBrace.Core
let sourceValues = [|for x in 0 .. 16 \rightarrow [|for y in 0 .. 2000 \rightarrow rand()|]|]
let cluster = ThespianCluster.InitOnCurrentMachine(4, ...)
   sourceValues
   |> CloudFlow.OfArray
   |> CloudFlow.map (fun x \rightarrow x |> Array.sum)
   |> CloudFlow.reduce (+)
   > cluster.Run
cluster.KillAllWorkers()
```

```
open MBrace.Core
let sourceValues = [|for x in 0 .. 16 \rightarrow [|for y in 0 .. 2000 \rightarrow rand()|]|]
let cluster = ThespianCluster.InitOnCurrentMachine(4, ...)
   sourceValues
   |> CloudFlow.OfArray
   |> CloudFlow.map (fun x \rightarrow x |> Array.sum)
   |> CloudFlow.reduce (+)
   > cluster.Run
cluster.KillAllWorkers()
```

```
open MBrace.Core
let sourceValues = [|for x in 0 .. 16 \rightarrow [|for y in 0 .. 2000 \rightarrow rand()|]|]
let cluster = ThespianCluster.InitOnCurrentMachine(4, ...)
   sourceValues
   |> CloudFlow.OfArray
   |> CloudFlow.map (fun x \rightarrow x |> Array.sum)
   |> CloudFlow.reduce (+)
   > cluster.Run
cluster.KillAllWorkers()
```

{m}brace

```
open MBrace.Core
let sourceValues = [|for x in 0 .. 16 \rightarrow [|for y in 0 .. 2000 \rightarrow rand()|]|]
let cluster = ThespianCluster.InitOnCurrentMachine(4, ...)
   sourceValues
   |> CloudFlow.OfArray
      CloudFlow.map (fun x \rightarrow x \mid > Array.sum)
      CloudFlow.reduce (+)
   > cluster.Run
cluster.KillAllWorkers()
```

{m}brace

```
open MBrace.Core
let sourceValues = [|for x in 0 .. 16 \rightarrow [|for y in 0 .. 2000 \rightarrow rand()|]|]
let cluster = ThespianCluster.InitOnCurrentMachine(4, ...)
   sourceValues
   |> CloudFlow.OfArray
   |> CloudFlow.map (fun x \rightarrow x |> Array.sum)
   |> CloudFlow.reduce (+)
   > cluster.Run
cluster.KillAllWorkers()
```

3 манипулирование данными

- repl & $f(\lambda)$
- Deedle
- Math.Net
- R Provider
- {m}brace



3 манипулирование данными

- repl & $f(\lambda)$
- Deedle
- Math.Net
- R Provider
- {m}brace
- Accord.Net Framework
- numl
- encog
- Hype
- ML ot MS
- AForge.Net
- ...



4 интеграция

это dotnet



ПОТИ

все прекрасно?



нет



нет, но можно



FSharp.Data RProvider {m}brace



python vs dotnet



почитать

- http://fsharp.org/
- https://fslab.org/
- http://fsharp.github.io/FSharp.Data/
- http://bluemountaincapital.github.io/Deedle/
- http://bluemountaincapital.github.io/FSharpRProvider/
- https://numerics.mathdotnet.com/
- https://tahahachana.github.io/XPlot/
- https://fslab.org/FSharp.Charting/
- http://accord-framework.net
- http://numl.net/

почитать

- http://www.heatonresearch.com/encog/
- https://azure.microsoft.com/ru-ru/services/machine-learning-studio/
- http://hypelib.github.io/Hype/
- http://mbrace.io/

почитать



Mastering .NET Machine Learning

F# for Quantitative Finance

An introductory guide to utilizing F# for quantitative finance leveraging the .NET platform

Johan Astborg

PACKT open source*

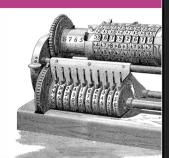
Master the art of machine learning with .NET and gain insight into real-world applications

Jamie Dixon

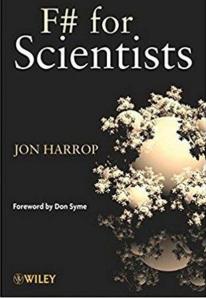
PACKT open source*

O'REILLY®

Analyzing and Visualizing Data with F#



Tomas Petricek



вопросы?

vk.com/oxffaa
akulyakov.artem@gmail.com
github.com/0xffaa